QUEUES WITH GALTON-WATSON-TYPE ARRIVALS

D. Fiems, J. Walraevens, H. Bruneel

SMACS Research Group, Department TELIN, Ghent University
Belgium
{df,jw,hb}@UGent.be

This paper presents the analysis of a discrete-time single server queueing system with a multi-type Galton-Watson arrival process with migration. It is shown that such a process allows to capture intricate correlation in the arrival process while the corresponding queueing analysis yields closed-form expressions for various moments of queue content and packet delay.

Keywords: Galton-Watson process, branching process, discrete-time queue.

1. INTRODUCTION

Input traffic at various nodes in packet switched telecommunication networks typically exhibits various levels of correlation. It is well known that input correlation significantly affects queueing performance and hence there is a continuing interest in analytically tractable queueing models which can accurately capture arrival correlation. There is a particular interest in Markovian arrival models, including models with a finite state space such as the discrete-time batch-Markovian arrival model [1, 2], or with a structured infinite state space such as the discrete autoregressive arrival models [3, 4] and the train and session arrival models [5, 6]. Queueing metrics for unstructured finite state space arrival models are not available in closed form. However, efficient algorithms are devised which yield the various performance measures in no time. In contrast, by imposing a structure on the state space of the arrival process, closed-form expressions for the various performance measures can be obtained.

In this paper, we propose a Markovian arrival process with a structured infinite state-space, the Galton-Watson arrival process. A discrete-time queueing system is analysed where the arrivals during the consecutive slots stem from a multi-type Galton-Watson branching processes with migration [7]. It is shown that such an arrival process exhibits intricate arrival correlation while closed form expressions for the probability generating functions of queue content and packet delay are obtained.

The remainder of the paper is organised as follows. In the next section, the arrival process and corresponding queueing model are introduced first. By a probability generating functions approach, we then obtain expressions for various performance measures. Our results are illustrated by some numerical examples in section 3. Finally, conclusions are drawn in section 4.

2. QUEUEING MODEL AND ANALYSIS

We consider a discrete-time queueing system; time is divided into fixed-length intervals or slots. During the consecutive slots, packets arrive at the queue, are stored in an infinite-capacity buffer and are served in order of arrival. Service times of the consecutive packets are fixed and equal to the slot length and packets cannot be served during their arrival slot.

Packet arrivals stem from a multi-type Galton-Watson process with migration; let K denote the number of types of this process. This branching process adheres the following recursion,

$$X_n^{(k)} = \sum_{i=1}^{K} \sum_{j=1}^{X_{n-1}^{(i)}} M_{j,n}^{(i,k)} + N_n^{(k)} \,. \tag{1}$$

Here $X_n^{(k)}$ denotes the number of arriving packets of type k during slot n, $M_{j,n}^{(i,k)}$ is the type k off-spring of the jth type i packet at slot $n-1$, and $N_n^{(k)}$ denotes the number of new type k packets at slot n. The total number of packet arrivals during slot n is then given by,

$$A_n = \sum_{k=1}^{K} X_n^{(k)} \tag{2}$$

The vectors $\left\{ \left[M_{j,n}^{(i,1)}, M_{j,n}^{(i,2)}, \dots M_{j,n}^{(i,K)} \right],\ j, n = 1, 2, \dots \right\}$ constitute a doubly indexed sequence of independent and identically distributed (iid) random vectors for all $i = 1, 2, \dots, K$. These random vectors are therefore completely characterised by the following vector-valued joint probability generating function,

$$\mathbf{M}(\mathbf{x}) = [M_i(\mathbf{x})]_{i=1,2,\dots,K} = \left[\mathrm{E}\left[\prod_{k=1}^{K} x_k^{M_{j,n}^{(i,k)}} \right] \right]_{i=1,2,\dots,K}, \tag{3}$$

with $\mathbf{x} = [x_1, x_2, \dots, x_K]$. Similarly, the vectors $\left\{ \left[N_n^{(1)}, N_n^{(2)}, \dots, N_n^{(K)} \right],\ n = 1, 2, \dots \right\}$ constitute a sequence of iid random vectors, characterised by the common joint probability generating function,

$$N(\mathbf{x}) = \prod_{k=1}^{K} \mathrm{E}\left[x_k^{N_n^{(k)}} \right] . \tag{4}$$

Finally let $\mu_{ik} = \mathrm{E}\left[M_{1,1}^{(i,k)} \right]$ denote the average type k offspring of a type i packet and let $\nu_k = \mathrm{E}\left[N_1^{(k)} \right]$ denote the mean number of new type k arrivals in a slot. Collecting these elements in the $K \times K$ matrix $\mathcal{M} = [\mu_{ik}]$ and in the column vector $\mathcal{V} = [\nu_k]$, the mean number of arrivals in a slot can be expressed as follows,

$$\rho = \mathrm{E}\left[A_1 \right] = \mathbf{e} \left(\mathcal{I} - \mathcal{M} \right)^{-1} \mathcal{V} \,, \tag{5}$$

with $\mathbf{e}$ a row vector of ones and with $\mathcal{I}$ the identity matrix. In the remainder, we assume $|\mathcal{M}| < 1$, $|\cdot|$ denoting the largest eigenvalue of its argument. Hence, the arrival load is finite.

With the notation of the arrival process established, we now focus on the queueing analysis. Let U_n denote the queue content at the beginning of slot n. The queue contents at the beginning of consecutive slots are then related as follows,

$$U_n = (U_{n-1} - 1)^+ + A_n \,. \tag{6}$$

Here $(\cdot)^+$ is the usual shorthand notation for $\max(\cdot, 0)$.

The state of the queueing system at slot boundary n is completely described in the Markovian sense by the vector of state variables $(U_n, X_n^{(1)}, \ldots, X_n^{(K)})$; see equations (1), (2) and (6). Therefore, let $P(\mathbf{x}, z)$ denote the joint probability generating function of this vector in steady state,

$$P(\mathbf{x}, z) = \lim_{n\to\infty} \mathrm{E}\left[\prod_{k=1}^{K} x_k^{X_n^{(k)}} z^{U_n}\right] . \tag{7}$$

It can be shown that the queueing system reaches steady state for $\rho < 1$ and $|\mathcal{M}| < 1$.

In view of equations (1), (2) and (6) and by standard z-transform techniques, it is found that $P(\mathbf{x}, z)$ satisfies the following functional equation,

$$P(\mathbf{x}, z) = P(\mathbf{M}(\mathbf{x}z), z)\frac{1}{z}N(\mathbf{x}z) - P(\mathbf{0}, 0)\frac{1-z}{z}N(\mathbf{x}z)\,, \tag{8}$$

with $\mathbf{0}$ a row vector of zeros. Here, we also used the fact that no packets arrive during slot $n-1$ when the queue is empty at the beginning of slot n. Let $\mathbf{Q}^{(i)}(\mathbf{x}, z)$ denote the row vector, recursively defined as follows,

$$\mathbf{Q}^{(i)}(\mathbf{x}, z) = \mathbf{M}(\mathbf{Q}^{(i-1)}(\mathbf{x}, z)z)\,, \quad \mathbf{Q}^{(0)}(\mathbf{x}, z) = \mathbf{x}\,, \tag{9}$$

for $i = 1, 2, \ldots$. Successive application of the functional equation (8) then yields,

$$P(\mathbf{x}, z) = P(\mathbf{0}, 0)(z-1)\sum_{j=0}^{\infty}\prod_{i=0}^{j}\frac{N(\mathbf{Q}^{(i)}(\mathbf{x}, z)z)}{z}\,. \tag{10}$$

Finally, the normalisation condition determines the remaining unknown $P(\mathbf{0}, 0) = 1-\rho$.

Clearly, the probability generating function of the queue content equals $U(z) = P(\mathbf{e}, z)$. Furthermore, given the probability generating function of the queue content, the probability generating function of the packet delay — the number of slots between the end of a packet's arrival and departure slot — is easily obtained by the distributional form of Little's result for discrete-time queues with single-slot service times [8],

$$D(z) = \frac{1}{\rho}(P(\mathbf{e}, z) - (1-\rho))\,. \tag{11}$$

The moment generating property of probability generating functions then yields the various moments of the queue content and packet delay.

3. NUMERICAL RESULTS

With the formulae at hand, we now study the mean delay of the Galton-Watson queueing system where the arrivals stem from a two-type Galton-Watson source with neither migration between the types ($\mu_{12} = \mu_{21} = 0$) nor correlation between the new arrivals of the different types. For this simplified arrival process, we introduce a simple parameter estimation procedure.

The autocorrelation function $\alpha(n)$ of this arrival process adheres,

$$\alpha(n) = \frac{\phi_1^2}{\phi^2}\,{\mu_{11}}^n + \frac{\phi_2^2}{\phi^2}\,{\mu_{22}}^n\,, \qquad \phi_i^2 = \frac{\theta_i^2(1-\mu_{ii}) + \nu_i\sigma_i^2}{(1+\mu_{ii})(1-\mu_{ii})^2} \tag{12}$$

Here ϕ_i^2 is the variance of the number of type-i arrivals in a slot and $\phi^2 = \phi_1^2 + \phi_2^2$ is the variance of the number of arrivals in a slot. Further, σ_i^2 and θ_i^2 denote the variances of $M_{1,1}^{(i,i)}$ and $N_1^{(i)}$, respectively.

To limit the number of parameters, assume that (i) $M_{1,1}^{(i,i)}$ is Bernoulli distributed and that (ii) $N_{1,1}^{(1)}$ and $N_{1,1}^{(2)}$ have the same index of dispersion (or variance-to-mean ratio), $\beta = \theta_1^2/\nu_1 = \theta_2^2/\nu_2$. In Fig. 1 the autocorrelation function of the arrival process is depicted for various parameter settings. The tangent $\alpha_0(n)$ in 0 and the asymptote $\alpha_\infty(n)$ are depicted as well in the logarithmic plot. These are given by,

$$\begin{aligned} \alpha_0(n) &= \exp(n(\kappa \ln \mu_{11} + (1-\kappa)\ln \mu_{22}))\,, \\ \alpha_\infty(n) &= \exp(n \ln \mu_{11} + \ln(\kappa))\,. \end{aligned} \tag{13}$$

with $\kappa = \phi_1^2/\phi^2$. Here, we assumed $\mu_{11} > \mu_{22}$, without loss of generality. Fig. 1 illustrates the versatility of the arrival model in capturing arrival correlation. Long- and short-time correlation can be adapted by modifying the decay rate of the tangent and the asymptote, respectively.

Clearly, the tangent and asymptote uniquely determine the parameters κ, μ_{11} and μ_{22} which in turn uniquely determine the autocorrelation function; see equations (12) and (13). Additionally fixing the total arrival load ρ and the variance of the number of arrivals in a slot ϕ^2, then uniquely determines the first and second order parameters of the arrival process, which are sufficient to obtain the mean delay. Hence, parameter estimation for a given trace reduces to (i) the estimation of the mean and variance of the number of arrivals in a slot and to (ii) estimating $\alpha_0(n)$ and $\alpha_\infty(n)$ for the empirical autocorrelation function of the trace. It can be shown that this procedure can always be applied for $\phi^2 \geq \rho$.

In Fig. 2, the mean delay is depicted versus the arrival load ρ for the various autocorrelation curves of Fig. 1 and for an index of dispersion $\phi^2/\rho = 2$. Given the load, the index of dispersion and the autocorrelation curve, all arrival process parameters are determined in accordance with the estimation procedure above. Fig. 2 clearly demonstrates the effect of arrival correlation on the various performance measures. In particular, a slow decay of long term correlation (cor3 and cor4) seriously affects the performance of the system. A similar observation also holds for short-term correlation (cor2 and cor4). However, the latter performance degradation is not as marked as the former

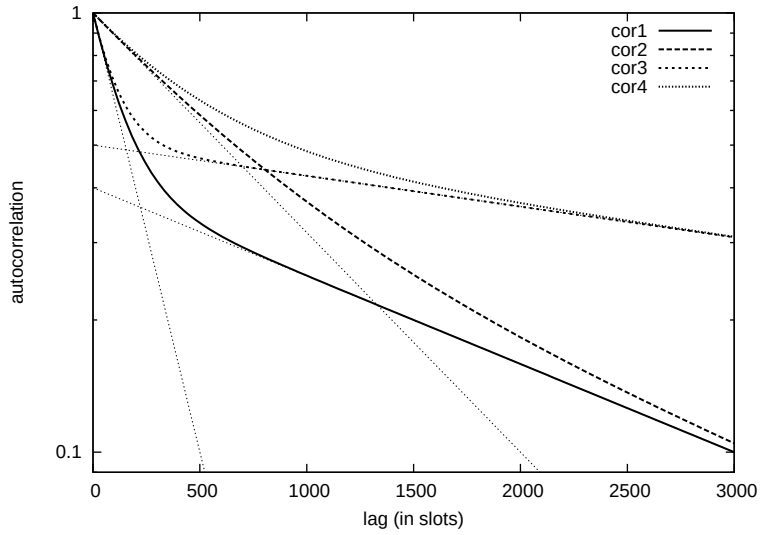

Fig. 1. Autocorrelation function of the arrival process for different parameter settings: $(\kappa, \mu_{11}, \mu_{22}) = (0.4, 0.999538, 0.992660)$ *for cor1,* $(0.4, 0.999538, 0.998391)$ *for cor2, for* $(0.5, 0.999839, 0.990991)$ *cor3 and* $(0.5, 0.999839, 0.997861)$ *for cor4.*

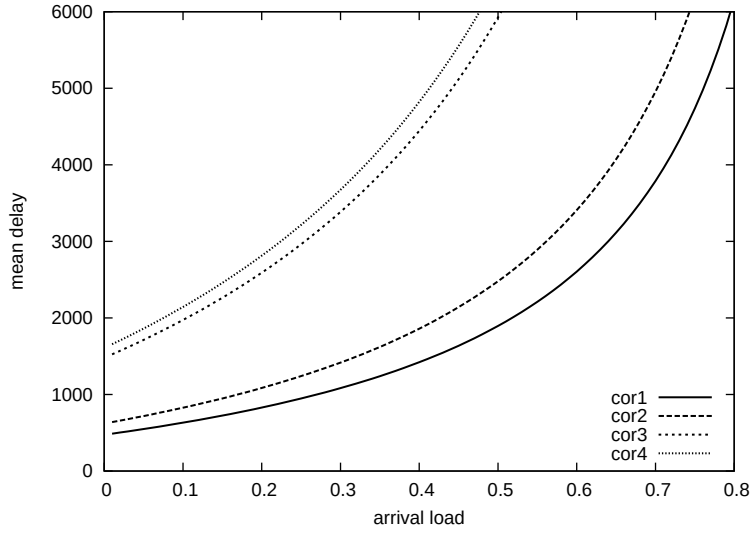

Fig. 2. Mean packet delay vs. the arrival rate for the parameter settings of Fig. 1 and for an index of dispersion $\phi^2/\rho = 2$.

4. CONCLUSIONS

In this paper, we provided closed-form expressions for the probability generating functions of the queue content and packet delay for the discrete-time queueing system with multi-type Galton-Watson arrivals. This queueing model is a versatile and tractable model for the analysis of buffers with arrival correlation. We showed that the parameters of a simplified Galton-Watson-type arrival process for a given empirical autocorrelation function are easily estimated. Given these parameters, the performance measures of interest are easily found. For a more accurate arrival characterisation, a more general Galton-Watson arrival model can be adopted.

ACKNOWLEDGEMENTS

The first two authors are postdoctoral fellows with the Research Foundation, Flanders (F.W.O.-Vlaanderen).

REFERENCES

1. *Blondia C., Casals O.* Statistical multiplexing of VBR sources: A matrix-analytic approach // Performance Evaluation. 1992. V. 16. P. 5–20.
2. *Herrmann C.* The complete analysis of the discrete time finite DBMAP/G/1/N queue // Performance Evaluation. 2001. V. 43. № 2–3. P. 95–121.
3. *Hwang G. U., Sohraby K.* On the exact analysis of a discrete-time queueing system with autoregressive inputs // Queueing Systems. 2003. V. 43. № 1–2. P. 29–41.
4. *Kamoun F.* The discrete-time queue with autoregressive inputs revisited // Queueing Systems. 2006. V. 54. № 3. P. 185–192.
5. *Wittevrongel S.* Discrete-time buffers with variable-length train arrivals // Electronics Letters. 1998. V. 34. № 18. P. 1719–1721.
6. *Hoflack L., De Vuyst S., Wittevrongel S., Bruneel H.* Analytic traffic model of web server // Electronics Letters. 2008. V. 44. № 1. P. 61–62.
7. *Athreya K. B., Ney P. B.* Branching Processes // Springer. 1972.
8. *Vinck B., Bruneel H.* Delay analysis for single server queues // Electronics Letters. 1996. V. 32. № 9. P. 802–803.